DOCUMENT RESUME

ED 167 069 IR 006 581

AUTHOR Gott, C. Deene

TITLE HIER-GRP: AA Computer Program for the Hierarchical Grouping of Regression Equations.

INSTITUTION Air Force Human Resources Lab., Brooks AFB, Texas.

REPORT NO AFHRL-TR-78-14

PUB DATE Jun 78

NOTE 28p.; Appendices C and D are not legible

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.

DESCRIPTORS Algorithms; *Cluster Analysis; *Computer Programs; *Multiple Regression Analysis; Statistical Analysis

ABSTRACT

This description of the technical details required for using the HIER-GRP computer program, which was develcped tc group or cluster regression equations in an iterative manner so as to minimize the overall loss of predictive efficiency at each iteration, contains a discussion of the basic algcrithm, an outline of the essential steps, specifications of the ccmputer system requirements, descriptions of necessary control cards, and explanaticns of the program output. Appendices include the mathematical formulas used, some mathematical background helpful for understanding the algorithm, sample output, and a complete source card listing. (Authcr/RAO)

HIER-GRP:

A Computer Program for the Hierarchical Grouping of Regression Equations

by

C. Deene Gott

Computational Sciences Division
Brooks Air Force Base, Texas 78235

June 1978

AFHRL-TR-78-14

TABLE OF CONTENTS

LIST OF TABLES

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFHRL-TR-78-14
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): HIER-GRP: A COMPUTER PROGRAM FOR THE HIERARCHICAL GROUPING OF REGRESSION EQUATIONS
5. TYPE OF REPORT & PERIOD COVERED: Final
6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): C. Deene Gott
8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Computational Sciences Division, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas 78235
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 62703F, 63230508
11. CONTROLLING OFFICE NAME AND ADDRESS: HQ Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base, Texas 78235
12. REPORT DATE: June 1978
13. NUMBER OF PAGES: 70
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
15. SECURITY CLASS. (of this report): Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

computer-based algorithm
computer program
grouping
hierarchical grouping
judgment analysis
policy capturing

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report describes the technical details required for using the HIER-GRP computer program as it is currently operational on the Univac 1108 computer system at the Computational Sciences Division, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas. HIER-GRP (or one of the earlier versions of the program) has been used extensively by the Air Force in the past, especially in conjunction with "policy-capturing applications," and many of those applications are referenced herein. The report contains a discussion of the basic algorithm, an outline of the essential steps, specifications of the computer system requirements, descriptions of necessary control cards, and explanations of the program output. Also, appendices are included that contain the mathematical formulas used, some mathematical background helpful for understanding the algorithm, sample output, and a complete source card listing.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

PREFACE

This research was completed under project 6323, Personnel Data Analyses; task 632305, Development of Analytic Methodology for Air Force Personnel Research Data.

In addition to the acknowledgments expressed in the introduction section of this report, the author wishes to give special credit to Mr. William S. Mathon for his numerous and valuable contributions to this project. Mathon performed the majority of the necessary programming tasks and prepared the basic text for Appendix B. Others who deserve mention include Mr. Larry K. Whitehead and Ms. Deana J. Olden for programming modifications and A1C Susan E. Tobey and Ms. Doris E. Black for technical editing. Finally, appreciation goes to Ms Dorothy M. Cobern and Ms. Laurel J. Betz for typing and proofreading the draft report.

HIER-GRP: A COMPUTER PROGRAM FOR THE HIERARCHICAL GROUPING OF REGRESSION EQUATIONS

I. INTRODUCTION

HIER-GRP, an acronym for hierarchical grouping, is a computer program which was developed for various Air Force research purposes at the Computational Sciences Division, Air Force Human Resources Laboratory, Brooks AFB, Texas. Given a starting set of k regression equations, each of which contains the same criterion and predictor variables, the basic objective of the HIER-GRP algorithm is to group or to cluster the equations in a stepwise or iterative manner so as to minimize the overall loss of predictive efficiency at each iteration. Initially there are k separate groups; i.e., each of the k equations is considered as a group by itself, and a measure of overall predictive efficiency is computed. At the first iteration all possible ways of combining any two of the equations from the total k equations are examined, and that combination providing the minimum loss of overall predictive efficiency is selected to form a "new group." Formation of the new group reduces the number of equations to k-1 for the start of the second iteration. The process continues until only one final group remains and is "hierarchical" in the sense that the pattern of the number of groups from start to finish is k, k-1, k-2, . . ., 1.

The mathematical theory upon which HIER-GRP is based is documented in an Air Force publication entitled *An Iterative Technique for Clustering Criteria Which Retains Optimum Predictive Efficiency* by Robert A. Bottenberg and Raymond E. Christal (3). Early developmental work was also accomplished by Joe H. Ward, Jr., (16), and some of the original programming was done by Daniel D. Rigney.

HIER-GRP or one of the earlier versions of the program has been used extensively by the Air Force in the past, especially in conjunction with "policy-capturing applications." Policy-capturing is a methodology composed of multiple linear regression analysis and hierarchical grouping procedures (1, 3, 4, 6, 7, 14, 16, 17, and 18). In this context, HIER-GRP was used in the development of the Weighted Airman Promotion System (WAPS) (10) and later in the reevaluation of WAPS (12 and 13). The program was also used in developing officer grade requirements (9), a promotion system for airman basics (2), a screening system for the Air Reserve Forces (8), and a senior NCO promotion system (11).

This report describes the technical details that are required for the use of the HIER-GRP program as it is currently operational on the Univac 1108 computer system at the Computational Sciences Division. The basic algorithm is first discussed, and the essential steps are outlined. Details of the computer system requirements and descriptions of necessary control cards are then presented. Next, the output of HIER-GRP is explained. Appendices are included that contain the mathematical formulas used in the program, some mathematical background helpful for understanding the algorithm, sample output, and a complete source card listing of the program.

Partly as a result of the research studies referenced above, requests for copies of the HIER-GRP computer program and associated documentation from different Air Force agencies, other governmental organizations, colleges, and universities have been numerous. Since 1969, approximately twenty copies of HIER-GRP have been provided to different requesters and implemented on a variety of different computer systems. One purpose of this report is to provide a document which can be used to satisfy any future requests for HIER-GRP.

II. BASIC ALGORITHM

This section describes the basic structure of the HIER-GRP algorithm. The reader is referred to Appendix A for computational formulas mentioned in the various steps and to Appendix B for more detailed mathematical considerations.

The basic steps of the HIER-GRP algorithm can be summarized as the following five phases: (a) data input and program termination, (b) computation of the overlap matrix, (c) determination of the order of clustering, (d) computation of the statistics for the initial k criteria, and (e) iteration to reduce the number of criteria. Each of these phases is described in the following steps. The steps are to be followed in numeric order unless indicated otherwise.

Steps 1–2. Data Input and Program Termination

1. Read "Problem Definition Card." This card defines k, the number of criteria or regression equations to be grouped and the number of standardized regression (beta) weights in each equation. If no Problem Definition Card is read, terminate the program.

2. Read in the number of cases, the criterion means and standard deviations, the standardized regression weights, the validities, and the predictor means and standard deviations for each equation. Assign each equation the identification numbers 1 through k, respectively, according to the order in which the equations were read.

Step 3. Computation of the Overlap Matrix.

3. Compute the overlap matrix A, where each element a_{ij} denotes the decrease in overall predictive efficiency if equation i is combined with equation j; for $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, k$, and $i \neq j$. The diagonal elements of A are undefined and the elements above the diagonal are symmetric with those elements below the diagonal.

Steps 4–8. Determination of the Order of Clustering

4. Set NGRPS, the index denoting the current number of groups, equal to k. Initially each criterion (equation) belongs to a separate cluster.

5. Considering all clusters present at the NGRPS stage, select two of the clusters denoted by i and j such that:

a. $a_{ij} \leq a_{\ell m}$ where ℓ and m are the identification numbers of any cluster present at the NGRPS stage and $\ell \neq m$, and

b. $i < j$. This can be accomplished by examining the elements above the diagonal of the overlap matrix and selecting the smallest element.

6. Form a new criterion cluster from the old clusters i and j identified in Step 5. Record the identifications of the two clusters i and j in the storage areas IU_{NGRPS} and JU_{NGRPS}, respectively. Assign the new cluster the identification number i.

7. Decrement NGRPS by 1. If NGRPS > 1, go to Step 8; otherwise proceed to Step 9.

8. Update the overlap matrix as follows. For each ℓ, $\ell \neq i$ of Step 6 where ℓ is the identification number of a criterion cluster present at the NGRPS stage, compute the decrease in overall predictive efficiency if equation ℓ is combined with equation i. Since NGRPS was reduced by 1 in Step 7, the dimension of the updated overlap matrix will be reduced by 1. Return to Step 5.

Step 9. Computation of the Statistics for the Initial k Criteria

9. Compute the squared multiple correlation coefficient for each of the initial k regression equations and, also, ORU_k, the overall squared multiple correlation coefficient obtained by considering a regression model with no grouping of initial equations.

Steps 10–15. Iteration to Reduce the Number of Criteria

10. Form an initial grouping of the k equations by assigning each equation to a group by itself. This is the "k groups" stage. Set NGRPS equal to k.

11. Form a new grouping of the k equations by following the grouping order established in Steps 4–8. This is accomplished by combining the groups identified by IU_{NGRPS} and JU_{NGRPS} and assigning the new group (criterion cluster) the identification number in IU_{NGRPS}.

12. Compute the least squares regression equation which can be used to predict the new group and decrement NGRPS by 1.

13. Print all statistics concerning the new grouping including:

a. the identification numbers of the two equations combined at this iteration,

b. An F value testing the difference between the prediction equations for the two clusters in (a),

c. An F value testing the difference between the k initial prediction equations and the smaller set of NGRPS equations (one for each cluster) used at the "NGRPS groups" stage, and

d. the overall squared multiple correlation coefficient obtained using the NGRPS equations at this stage.

14. Print a summary of all groups (clusters) present at the NGRPS stage. Also, print the prediction equation for the new group (including standardized and raw score weights).

15. If NGRPS > 1, loop back to Step 11; otherwise, return to Step 1 and begin the next problem.

III. DESCRIPTIONS OF THE HIER-GRP PROGRAM

Systems Requirements

The HIER-GRP program is composed of seven routines–the main or driver routine and six subroutines. The entire program, with the exception of the Univac Assembly Language subroutine START, is written in FORTRAN IV. The assembly subroutine START is called once at the beginning of the driver routine and is never called again. Its only function is to reset the margin control on the Univac 1108 printer.

The Univac version of FORTRAN has a special statement, the Parameter statement, which is used in the driver routine and which may not be available on other computers. The Parameter statement is used to define the dimensions of arrays at compilation time. The Parameter statement can be removed if each array is dimensioned to its required size.

The complete HIER-GRP program requires approximately 10,000 36-bit words of core storage in addition to the number of words required for arrays. If P is the number of predictors and E is the number of equations, then the amount of storage required for arrays is $12E+3P+[2\cdot E\cdot P]+[E\cdot(E-1)/2]+14$. For example, if P = 50 and E = 50, then 6,989 words of storage are required for arrays.

There are a total of 1,121 cards in the HIER-GRP program deck. Of these, only 601 are source language cards and the remainder are comments cards. The number of cards and the intrinsic system routines required in each of the seven routines which make up HIER-GRP are listed in Table 1.

Table 1. **Characteristics of the HIER-GRP Routines**

Program Name	Source Language	Number of Source Language Cards	Number of Comment Cards	Intrinsic System Routines Required
DRIVER (MAIN)	FORTRAN IV	100	311	None
START	ASSEMBLY	7	0	None
OVRLP	FORTRAN IV	36	36	None
GROUP	FORTRAN IV	76	48	None
STAGE	FORTRAN IV	81	42	None
PRINTG	FORTRAN IV	218	82	SQRT
PLEVEL	FORTRAN IV	83	1	ATAN, SQRT, ALOG, EXP, SIN

Data Requirements

A HIER-GRP user must supply the following data for each regression equation:

1. The number of cases (N) which were used to compute the equation
2. The criterion mean and standard deviation (SD)
3. The standardized regression weights
4. The validity coefficients (correlations of predictor or independent variables with the criterion or dependent variable)
5. The predictor means and standard deviations.

The computational formulas developed by Bottenberg and Christal (3) and used within the program assume that the predictor sums-of-squares and cross-products matrices are proportional, i.e., that the ratios of the corresponding elements of the sums-of-squares and cross-products matrices for any two equations to be clustered are equal to the ratio of the corresponding numbers of the cases within each equation. This assumption of proportionality is discussed in detail by Bottenberg and Christal (1961, see pages 8 through 11) and also addressed in Appendix B (see equation 9b) of this report. In practice this assumption is met by selecting items (1) and (5) of the previous paragraph to be identical for each equation.

Run-Stream Organization

The following card sequence is required to use the HIER-GRP program as it is operational on a Univac 1108 computer:

Order	Card Type
1.	@RUN
2.	@XQT T*T.HIER-GRP
3.	Problem Definition Card
4.	Header Card(s)
5.	Format Card for Equation Ns
6.	Data Card(s) – Equation Ns
7.	Format Card for Criterion Means and SDs
8.	Data Card(s) – Criterion Means and SDs
9.	Format Card for Beta Weights
10.	Data Card(s) – Beta Weights
11.	Format Card for Validities

12. Data Card(s) – Validities
13. Format Card for Predictor Means and SDs
14. Data Card(s) – Predictor Means and SDs
15. The sequence of cards 3–14 is required for each run. As many problems as desired may be run by stacking one problem after another.
16. Blank Card to Terminate Run
17. @FIN

The Univac 1108 System Cards (1, 2, and 17) are described in the Univac Exec 8 Reference Manual (15). Descriptions of cards 3–16 are presented in the next section. See Appendix C for sample run-stream and sample control cards.

Control Cards

Problem Definition Card

Card Columns	FORTRAN Format		Description
1–3	I3	NEQS,	the number of criteria (systems, regression equations) in this problem. NEQS must be less than or equal to 50.
4–6	I3	NPREDS,	the number of beta weights (standardized regression weights) in each equation. NPREDS must be less than or equal to 100.
7	I1	IOPT,	the grouping (clustering) option desired. Normally a "6" is specified which causes the grouping to be done based on the iterative technique developed by Bottenberg and Christal (3). Other options are included in the program and comments cards, but are for future developmental purposes only.
8	I1	NHDRS,	the number of header (label, title) cards that follow this control card. Header cards can be omitted (NHDRS = 0) or up to 9 cards may be specified.
9	I1	IREAD,	the data read option. IREAD = 0 means read the beta weights and validities NPREDS items at a time. IREAD = 1 means read them NEQS*NPREDS items at a time. IREAD allows flexibility in the format of input data. However, IREAD is normally set equal to zero.
10–80			These card columns are not read.

Header Cards

Each header card will be printed only once at the beginning of the grouping report. Exactly NHDRS header cards must be present.

Format and Data Cards

1. *Format Card for Equation Ns.* This card supplies the FORTRAN variable format by which the number of cases used in the computation of each equation is to be read. Only the F and X editing codes are permitted.

2. *Data Card(s) Equation Ns.* These cards are read according to the previous format card. The number of cards required depends on the format specifications.

3. *Format Card for Criterion Means and SDs.* This card provides the FORTRAN variable format by which the criterion mean and standard deviation for each equation are to be read. Only the F and X editing codes are permitted.

4. *Data Card(s) Criterion Means and SDs.* These cards are read according to the previous format card. The number of cards required depends on the format specifications.

5. *Format Card for Beta Weights.* This card supplies the FORTRAN variable format by which the beta weights (NPREDS weights per equation) are to be read. Only the F and X editing codes are permitted.

6. *Data Card(s) Beta Weights.* These cards are read according to the previous format card. Exactly NEQS sets of cards are required if IREAD = 0. The first set contains the beta weights for equation 1, the second set contains the beta weights for equation 2, and so on. The number of cards within each set depends on the format specifications.

7. *Format Card for Validities.* This card provides the FORTRAN variable format by which the validity coefficients for each equation are read. Only the F and X editing codes are permitted.

8. *Data Card(s) Validities.* These cards are read according to the previous format card. Exactly NEQS sets of cards are required if IREAD = 0. The first set contains the validities for equation 1, the second set contains the validities for equation 2, and so on. The number of cards within each set depends on the format specifications.

9. *Format Card for Predictor Means and SDs.* This card supplies the FORTRAN variable format by which the predictor means and standard deviations for each equation are to be read. Only the F and X editing codes are permitted.

10. *Data Card(s) Predictor Means and SDs.* These cards are read according to the previous format card. The number of cards required depends on the format specifications.

Output

The printed output of HIER GRP is divided into five parts — the monogram and version date, the control card parameters, the problem header label, the format and input data cards, and the criterion grouping results. Each of these divisions is described in the following paragraphs. Refer to Appendix C for sample output.

Monogram and Version Date

The program title "Hierarchical Grouping Program HIER GRP," the AFHRL monogram, and the program version date are printed at the beginning of each problem. The program version date is the last time the program was updated or modified.

Control Card Parameters

The parameters specified on the Problem Definition card are printed under the heading CONTROL CARD PARAMETERS. This includes the number of regression equations (criteria), the number of beta weights in each equation, the grouping option desired, and the number of header cards for this problem.

Problem Header Label

The problem header label, if header cards were specified on the Problem Definition Card, is printed under the heading PROBLEM HEADER LABEL.

Format and Input Data Cards

All format cards and all input data are printed under the heading FORMAT CARDS AND INPUT DATA. First, the format statements used to read the number of cases and the criterion means and standard deviations for each equation are printed. A table listing the equation numbers, the number of cases, the criterion means, and the criterion standard deviations is printed next. Third, the format statement used to read the beta weights and a table listing the equation number and the beta weights (15 per line) for each equation are printed. Fourth, the format statement used to read the validity coefficients, and a table listing the equation number and the validities (15 per line) for each equation are printed. Finally, the format statement used to read the predictor means and standard deviations and a table listing the predictor variable number and predictor means and standard deviations (one each per line) are printed.

Criterion Grouping Results

The results of the clustering process are printed under the heading HIERARCHICAL GROUPING RESULTS. The output in this division can be separated into three parts – the grouping option description, the R-square (RSQ) summary for the NEQS initial criteria, and the results of each iteration. Each of these sections is described as follows.

1. *Grouping Option Description.* The grouping option and a verbal description of the grouping option specified on the Problem Definition Card are printed.

2. *RSQ Summary for the NEQS Initial Criteria.* The number, NEQS, of initial criteria; the overall RSQ, ORU_{NEQS}, achieved by using the beta weights specified on the input data cards; and a table listing the equation number and the RSQ for each equation are printed.

3. *Results of Each Iteration.* The statistics and tables printed at each iteration, i.e., the information printed below each row of asterisks is listed as the following in Table 2.

13

Table 2. **Output for Each Iteration**

Computer Output Label	Meaning
Stage = ℓ	ℓ is the number of criterion clusters present at the end of this iteration.
OVERALL/RSQ = ORU_{ℓ}	This is the RSQ obtained by using ℓ equations (one for each criterion cluster present at this stage) to predict the NEQS initial criteria.
SYSTEMS GROUPING THIS STAGE Table	
SYS IDENT	The identification (ID) numbers of the two criterion clusters combined at this iteration.
NO. MEMBERS	The number of members in each criterion cluster. The ID numbers of the members of each cluster can be obtained by referring to the summary roster for stage $\ell+1$.
NO. CASES	The number of cases used in the computation of the prediction equation for each criterion cluster. This number is the sum of the number of cases used in the prediction equation for each member of the cluster.
RSQ	The squared multiple correlation coefficient which is obtained by predicting each criterion within a cluster from the same compromise regression equation.
DECISION VALUE	The loss associated with replacement of the two clusters combined at this stage.
F-TEST FOR THE EQUALITY OF REGRESSION PARAMETERS FOR SYS'S COMBINED AT THIS STAGE Table	This table outlines a test of the hypothesis that the prediction equations for the two criterion clusters combined at this stage are identical. Equivalently, it is a test of the loss in predictive efficiency when each criterion within the two clusters combined at this stage are predicted from the same compromise equation.
CHANGE FROM $\ell+1$ SYSTEMS	
RSQ = $ORU_{\ell+1} - ORU_{\ell}$	The decrease in OVERALL RSQ from stage $\ell+1$.
DF = NPREDS+1	The decrease in the number of parameters estimated from stage $\ell+1$.
RESIDUAL	
RSQ = $1-ORU_{\ell+1}$	The proportion of the criterion variance attributable to error at stage $\ell+1$.
DF = N−(ℓ+1)(NPREDS+1)	The total number of cases less the number parameters estimated at stage $\ell+1$. Equivalently, DF is the number of degrees of freedom associated with the error portion of the criterion variance at stage $\ell+1$.
FSTAT = $[(ORU_{\ell+1}-ORU_{\ell})/(NPREDS+1)] / [(1-ORU_{\ell+k})/(N-(\ell+1)(NPREDS+1))]$	The F statistic testing the hypothesis described in the preceding paragraph (FOR SYS'S COMBINED AT THIS STAGE)

Computer Output Label	Meaning
SIG LVL	The probability that a value of the F statistic greater than FSTAT would occur by chance. A value of SIG LVL equal to α means that if the hypothesis being tested is true, then a value of the F statistic greater than FSTAT would have occurred 100α percent of the time by chance. Therefore, small values of α tend to reject the hypothesis being tested.
F-TEST FOR THE EQUALITY OF REGRESSION PARAMETERS FOR SYS'S COMBINED UP TO THIS STAGE Table	This table outlines a test of the hypothesis that the prediction equations for all members of criterion cluster number 1 are identical, the prediction equations for all members of criterion cluster 2 are identical, and so on for the ℓ criterion clusters present at the end of this iteration. Equivalently, this tests the loss in predictive efficiency when ℓ equations (one for each criterion cluster) are used to predict the NEQS initial criteria instead of the original NEQS equations.
CHANGE FROM NEQS SYSTEMS	
$RSQ = ORU_{NEQS} - ORU_{\ell}$	The decrease in OVERALL RSQ from stage NEQS.
$DF = (NEQS-\ell)(NPREDS+1)$	The decrease in the number of parameters estimated from stage NEQS.
RESIDUAL	
$RSQ = 1 - ORU_{NEQS}$	The proportion of the criterion variance attributable to error at stage NEQS.
$DF = N-(NEQS)(NPREDS+1)$	The total number of cases less the number of parameters estimated at stage NEQS. Equivalently, DF is the number of degrees of freedom associated with the error portion of the criterion variance at stage NEQS.
$FSTAT = [(ORU_{NEQS} - ORU_{\ell})/(NEQS-\ell)(NPREDS+1)] / [(1-ORU_{NEQS})/(N-(NEQS)(NPREDS+1))]$	The F statistic testing the hypothesis described in the preceding paragraph (FOR SYS'S COMBINED UP TO THIS STAGE)
SIG LVL	The probability that a value of the F statistic greater than FSTAT would occur by chance. A value of SIG LVL equal to α means that if the hypothesis being tested is true, then a value of the F statistic greater than FSTAT would have occurred 100α percent of the time by chance. Therefore, small values of α tend to reject the hypothesis being tested.
SYSTEMS SUMMARY ROSTER Table	The summary roster is a listing of all the criterion clusters present at the end of the current iteration. The members and the RSQ for each cluster are also printed. In addition, the prediction equation and the system mean and standard deviation for the new criterion cluster formed at the present iteration are printed. The compromise equation for each criterion cluster present at a given iteration can be obtained by referring to the summary roster for the stage at which the cluster was formed.

Table 2. (*Continued*)

Computer Output Label	Meaning
STAGE IDENT	The stage at which each criterion cluster was formed.
SYS LOSS	The contribution of each criterion cluster to the decrease in OVERALL RSQ from stage NEQS. Equivalently, this is the amount by which the OVERALL RSQ would increase if each of the criteria within this cluster were predicted from their individual regression equations rather than from the compromise equation for the cluster.
NO. MEMBERS	The number of criteria within each criterion cluster. The ID numbers of the members of each cluster are listed under the headings SYS IDENT and IDENTIFICATION OF OTHER MEMBERS in this table.
RSQ	The squared multiple correlation coefficient which is obtained by predicting each criterion within a cluster from the same compromise regression equation.
NO. CASES	The number of cases used in the computation of the compromise equation for a criterion cluster. This number is the sum of the number of cases used to compute the regression equation for each criterion within the cluster.
SYS IDENT	The ID number of a criterion cluster. This is also the smallest ID number of the criteria within this cluster.
IDENTIFICATION OF OTHER MEMBERS	The ID numbers of the remaining criteria within a cluster.
NEW SYS CRITERION MEAN	The criterion mean for the cluster formed at this iteration.
NEW SYS CRITERION SD	The criterion standard deviation for the cluster formed at this iteration.
BETA WEIGHTS FOR THE NEW SYSTEM S	The values (10 per line) of the least squares standardized regression coefficients for the regression equation which is the best single predictor for all the criteria in the new cluster where S is the ID number of the new cluster. Equivalently, these are the beta weights which would be obtained by pooling the observations for all the criteria in the new cluster and computing the regression of the pooled criteria on the NPREDS predictor variables.
RAW SCORE WEIGHTS FOR THE NEW SYSTEM S	The values (5 per line) of the raw score weights for the regression equation which is the best single predictor for all the criteria in the new cluster S.
REGRESSION CONSTANT	The regression constant for the regression equation which is the best single predictor of all the criteria in the new cluster.
Y SINGLE MEMBER SYSTEMS	A list of the identification numbers of the "Y" single criteria which have not been combined with any system up to this stage.

REFERENCES

1. **Anderberg, M.R.** *Cluster analysis for applications.* OAS-TR-72-1, AD-738-301. Kirtland AFB, NM: Office of the Assistant for Study Support, January, 1972.

2. **Black, D.E.** *Development of the E-2 weighted airman promotion system.* AFHRL-TR-73-3, AD-767 195. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, April 1973.

3. **Bottenberg, R.A., & Christal, R.E.** *An interative technique for clustering criteria which retains optimum predictive efficiency.* WADD-TN-61-30, AD-261 615. Lackland AFB, TX: Personnel Laboratory, Wright Air Development Division, March 1961. Also, *Journal of Experimental Education*, Summer 1968, **36**(4), pp. 28–34.

4. **Bottenberg, R.A., & Ward, J.H., Jr.** *Applied multiple linear regression.* PRL-TDR-63-6, AD-413 128. Lackland AFB, TX: 6570th Personnel Research Laboratory, Aerospace Medical Division, March 1963.

5. **Brown, B.** Simple comparisons of simultaneous regression lines. *Biometrics*, 1970, **26**, pp. 143–144.

6. **Christal, R.E.** *JAN: A technique for analyzing group judgment.* PRL-TDR-63-3, AD-403 813. Lackland AFB, TX: 6570th Personnel Research Laboratory, Aerospace Medical Division, February 1963.

7. **Christal, R.E.** *Selecting a harem–and other applications of the policy-capturing model.* PRL-TDR-67-1, AD-658 025. Lackland AFB, TX: Personnel Research Laboratory, Aerospace Medical Division, March 1967.

8. **Gott, C.D.** *Development of the weighted airman screening system for the air reserve forces.* AFHRL-TR-74-18, AD-781 747. Lackland AFB, TX: Computational Sciences Division, Air Force Human Resources Laboratory, March 1974.

9. **Hazel, J.T., Christal, R.E., & Hoggatt, R.S.** *Officer grade requirements project: IV. Development and validation of a policy equation to predict criterion board ratings.* PRL-TR-66-16, AD-659 125. Lackland AFB, TX: Personnel Research Laboratory, Aerospace Medical Division, November 1966.

10. **Koplyay, J.B.** *Field test of the weighted airman promotion system: Phase I. Analysis of the promotion board component in the weighted factors system.* AFHRL-TR-69-101, AD-689 751. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, April 1969.

11. **Koplyay, J.B., Albert, W.G., & Black, D.E.** *Development of a senior NCO promotion system.* AFHRL-TR-76-48, AD-A030 607. Lackland AFB, TX: Computational Sciences Division, Air Force Human Resources Laboratory, July 1976.

12. **Koplyay, J.B., & Gott, C.D.** *Reevaluation of the operational weighted airman promotion system for grades E-5 through E-7.* AFHRL-TR-73-25, For Official Use Only. Lackland AFB, TX: Computational Sciences Division, Air Force Human Resources Laboratory, November 1973.

13. **Koplyay, J.B., & Gott, C.D.** *Revalidation of the factors which comprise the E-5/E-7 weighted airman promotion system (WAPS).* AFHRL-TR-77-80, For Official Use Only. Brooks AFB, TX: Computational Sciences Division, Air Force Human Resources Laboratory, December 1977.

14. **Martin, F.B., & Zyskind, G.** On Combinability of Information from Uncorrelated Linear Models by Simple Weighting. *Annals of Mathematical Statistics*, Aug-Dec 1966, 37, pp. 1338–1347.

15. Sperry Rand Corporation. Univac 1100 Series Operating System, Programmer Reference, UP–4144 Rev. 3, 1974.

16. **Ward, J.H., Jr.** *Hierarchical Grouping to Maximize Payoff.* WADD-TN-61-29, AD-261 750. Lackland AFB, TX: Personnel Laboratory, Wright Air Development Division, March 1961.

17. **Welch, B.L.** Some problems in the analysis of regression among K samples of two variables. *Biometrika*, 1935, **27**, pp. 145–160.

18. **Wilson, J.W., & Carry, L.R.** Homogeneity of regression – its rationale, computation, and use. *American Educational Research Journal*, 1969, **6**, pp. 80–90.

BIBLIOGRAPHY

1. **Carter, A.H.** The estimation and comparison of residual regressions where there are two or more related sets of observations. *Biometrika*, 1949, **36**, pp. 26–46.

2. **Chaud, U.** Distributions related to comparison of two means and regression coefficients. *Annals of Mathematical Statistics*, 1950, **21**, pp. 507-521.

3. **Chipman, J.S., & Rao, M.M.** The treatment of linear restrictions in regression analysis. *Econometrica*, Jan–Apr 1964, 132(1–2), pp. 198–209.

4. **Fraser, A.S.** The Behrens-Fisher problem for regression coefficients. *Annals of Mathematical Statistics*, 1953, **24**, pp. 390–402.

5. **Geeslin, W.E.** Comment on homogeneity of regression. *American Educational Research Journal*, 1970, 7, pp. 636–638.

6. **Kendall, M.G., & Stuart, A.** *The advanced theory of statistics*, Vol. 2, *Inference and relationship* (Vol. 2), New York: Hafner, 1961.

7. **Kullback, S., & Rosenblatt, H.M.** On the analysis of multiple regression in K categories. *Biometrika*, 1957, **44**, pp. 67–83.

8. **Rao, C.R.** *Linear statistical inference and its applications.* New York: Wiley, 1965.

9. **Robson, D.S., & Atkinson, G.F.** Individual degrees of freedom for testing homogeneity of regression coefficients in a one-way analysis of covariance. *Biometrics*, 1960, **16**, pp. 593–605.

10. **Theil, H.** *Principles of econometrics.* New York: Wiley, 1970.

11. **Williams, E.G.** *Regression analysis.* New York: Wiley, 1959.

APPENDIX A: NOTATION AND COMPUTATIONAL FORMULAS

The transpose of the associated matrix.

k, The initial number of criteria.

p, The number of variables.

n_i, The number of cases used in the computation of the regression equation for criterion i.

m_i, The mean for criterion i.

σ_i^2, The variance for criterion i.

$\hat{\alpha}_i$, The constant term in the regression equation for criterion i.

$\hat{b}_i$, The pxl vector of regression weights for criterion i.

$\hat{\beta}_i$, The pxl vector of standard regression weights for criterion i.

c_i, The pxl vector of validities (intercorrelations between the criterion and the p independent variables) for criterion i.

N, The total number of cases $N = n_1+n_2+\ldots+n_k$

m_o The pooled criterion mean $Nm_o = n_1m_1+n_2m_2+\ldots+n_km_k$

σ_o^2, The pooled criterion variance

$$N\sigma_o^2 = n_1(\sigma_1^2+m_1^2)+\ldots+n_k(\sigma_k^2+m_k^2) - Nm_o^2$$

g_I, The number of criteria in cluster I.

I, The set of criteria in cluster I. $I = \{i_1, i_2, \ldots, i_{g_I}\}$. In the succeeding definitions, let I be the union of clusters J and L, $J \cup L$.

N_I, The number of cases used in the computation of the composite equation for cluster I.

$$N_I = \sum_{i\epsilon I} n_i = N_J+N_L$$

M_I, The criterion mean for cluster I.

$$N_IM_I = \sum_{i\epsilon I} n_im_i = N_JM_J+N_LM_L$$

σ_I^2, The criterion variance for cluster I.

$$N_I\sigma_I^2 = \sum_{i\epsilon I} n_i(\sigma_i^2+m_i^2) - N_IM_I^2 = N_J(\sigma_J^2+M_J^2)+N_L(\sigma_L^2+M_L^2) - N_IM_I^2$$

$\hat{\alpha}_I$, The constant term in the regression equation for cluster I.

$$N_I\hat{\alpha}_I = \sum_{i\epsilon I} n_i\hat{\alpha}_i = N_J\hat{\alpha}_J+N_L\hat{\alpha}_L$$

$\hat{b}_I$, The pxl vector of regression weights for cluster I.

$$N_I\hat{b}_I = \sum_{i\epsilon I} n_i\hat{b}_i = N_J\hat{b}_J + N_L\hat{b}_L$$

$\hat{\beta}_I$, The pxl vector of standard regression weights for cluster I.

$$N_I \sigma_I \beta_I = \sum_{i \epsilon I} n_i \sigma_i \hat{\beta}_i = N_J \sigma_J \beta_J + N_L \sigma_L \hat{\beta}_L$$

c_I, The pxl vector of validities for cluster I.

$$N_I \sigma_I c_I = \sum_{i \epsilon I} n_i \sigma_i c_i = N_J \sigma_J c_J + N_L \sigma_L c_L$$

R_i^2, The squared multiple correlation coefficient for the regression on criterion i.

$$R_i^2 = \hat{\beta}_i' c_i$$

R_I^2, The squared multiple correlation coefficient for the regression on cluster I.

$$R_I^2 = \hat{\beta}_I' c_I = \frac{1}{N_I^2 \sigma_I^2} \left[N_J^2 \sigma_J^2 R_J^2 + N_L^2 \sigma_L^2 R_L^2 + N_J N_L \sigma_J \sigma_L (\hat{\beta}_J' c_L + \hat{\beta}_L' c_J) \right]$$

G_s, The set of s criterion clusters present at the s cluster stage.

$$G_s = \{ I_1, I_2, \ldots, I_s \} .$$

${}_sR^2$, The squared multiple correlation coefficient for the criterion grouping, G_s, at the s cluster stage.

$$N\sigma_o^2 \, {}_sR^2 = \sum_{I \epsilon G_s} N_I (\sigma_I^2 R_I^2 + M_I^2) - Nm_o^2 .$$

Let $G_s = \{ J, L, K_3, \ldots, K_s \}$ and

$G_{s-1} = \{ J \cup L, K_3, \ldots, K_s \}$ then

$${}_sR^2 - {}_{s-1}R^2 = \frac{N_J N_L}{N\sigma_o^2 (N_J + N_L)} \left[\sigma_J^2 R_J^2 + \sigma_L^2 R_L^2 + (M_J - M_L)^2 - \sigma_J \sigma_L (\hat{\beta}_J' c_L + \hat{\beta}_L' c_J) \right]$$

20

APPENDIX B: MATHEMATICAL BACKGROUND

Mathematical Model for the Clustering Algorithm

Suppose that a set of p independent variables, $v' = (v_1, \ldots, v_p)$, are linearly related to the expected values of each of k criteria, $Y_1, \ldots Y_k$; that is,

(1) $E(Y_i|v) = v'b_i + \alpha_i$ for $i=1, \ldots, k$,

where b_i is a pxl vector of unknown population parameters and α_i is an unknown population constant. Let y_i be an $n_i$xl vector of independent observations on criterion Y_i, let X_i be an $n_i$xp matrix of observations on the set of p independent variables v, where the j–th element of y_i corresponds to the j–th row of X_i, and let u_i be an $n_i$xl vector in which each element is 1. Then from (1),

(1a) $E(y_i|X_i) = X_i b_i + u_i \alpha_i$ for $i=1, \ldots, k$.

Let $N = n_1 + \ldots + n_k$; let $Y' = [y_1', \ldots, y_k']$, the 1xN vector obtained by pooling all the criterion observations; let

$$X = \begin{bmatrix} u_1 X_1 & 0 & 0 & . & . & . & 0 \\ 0 & u_2 X_2 & 0 & . & . & . & 0 \\ 0 & . & . & . & . & . & u_k X_k \end{bmatrix}$$

the Nxk(p+1) block diagonal matrix obtained by placing the n_ix(p+1) matrix of observations $[u_i X_i]$ in the i–th block diagonal position, and let $b' = [\alpha_1 b_1' \ldots \alpha_k b_k']$; the k(p+1) vector of unknown parameters. Under the assumption that the criterion observations are independent and have common variance, the mathematical model for the clustering algorithm is

(1b) $E(Y|X) = Xb$ with $D(Y|X) = \sigma^2 I$,

where $D(Y|X)$ is the dispersion matrix of the criterion observations, σ^2 is the common variance, and I is the NxN identity matrix.

Minimum Variance Unbiased Estimation and Hypothesis Testing

The k(p+1)xl vector b of unknown parameters in (1b) correspond to the k equations in (1a). The minimum variance unbiased estimates (mvue), $\hat{\alpha}_i$ and $\hat{b}_i$, of α_i and b_i are obtained from (1b) by the method of least squares, where

$$(2) \quad \begin{aligned} \hat{b}_i &= \left[X_i'X_i - \frac{1}{n_i} X_i'u_iu_i'X_i\right]^{-1} \left[X_i'y_i - \frac{1}{n_i} X_i'u_iu_i'y_i\right] \\ \hat{\alpha}_i &= \frac{1}{n_i} u_i'y_i - \frac{1}{n_i} u_i'X_i\hat{b}_i \end{aligned} \qquad \text{for } i=1, \ldots, k.$$

These are the estimates that would be obtained by the method of least squares from the k separate models

(3) $E(y_i|X_i) = X_i b_i + u_i \alpha_i$ with $D(y_i|X_i) = \sigma^2 I$ for $i=1, \ldots, k$

where the error variance, σ^2, is the same for each model. It might be that some or all of the equations in (1) are identical. The technique of homogeneity of regression can be used to test the equality of vectors of regression parameters across several criteria. Chipman and Rao (1964) and Theil (1970) have developed methods for obtaining mvue under general linear restrictions and for testing general linear hypotheses. Rao (1965, pp 189–190) shows that in the case

(4) $E(Y|X) = Xb$ with $D(Y|X) = \sigma^2 I$,

where X is nxs of rank s and b is sxl, the mvue, $\hat{b}_\Psi$ for b under the linear restriction

(4a) $\Psi b = 0$ is

(4b) $\hat{b}_\Psi = B(B'X'XB)^{-1}B'X'Y$

21

where Ψ is rxs of rank r, B is sx(s-r) of rank (s-r), and $\Psi B = 0$. Rao obtains this result by introducing the general solution, $B\theta$, where θ is an (s-r)x1 vector of new parameters, of (4a) into (4) to obtain the model

(5) $E(Y|X) = XB\theta$ with $D(Y|X) = \sigma^2 I$

and no restrictions on θ. The mvue, $\hat{B\theta}$, of $B\theta$ is $B\hat{\theta}$ (see Rao, 1965, pp. 181–182), where $\hat{\theta}$ is the mvue of θ in (5). If, in addition to (4), Y has a multivariate normal distribution, then Chipman and Rao develop an expression for an unbiased critical region of size θ for the following hypothesis:

(6) $\Psi_1 b = 0$ given that $\Psi_0 b = 0$

where Ψ_1 is $r_1$xs of rank r_1, Ψ_0 is $r_0$xs of rank r_0, and $\Psi' = [\Psi_0' \Psi_1']$ is sx (r_0+r_1) of rank (r_0+r_1). The expression for the unbiased critical region of size θ is

(7) $$\left\{F \middle| F = \left(\frac{n-s+r_0}{r_1}\right)\left(\frac{EXSS}{ESSH}\right) = \left(\frac{n-s+r_0}{r_1}\right)\left(\frac{R^2_{\Psi_0} - R^2_{\Psi}}{1-R^2_{\Psi_0}}\right) > F_\theta\left(r_1, n-s+r_0\right)\right\},$$

where $F_\theta(r_1, n-s+r_0)$ is the upper 100 $(1-\theta)$% point of the central F distribution with r_1 and $n-s+r_0$ degrees of freedom, and

$ESSH = (Y-X\hat{b}_{\Psi_0})'(Y-X\hat{b}_{\Psi_0})$,

$EXSS = (Y-X\hat{b}_{\Psi})'(Y-X\hat{b}_{\Psi}) - ESSH$,

$\hat{b}_{\Psi_0}$ is the mvue of b under the restriction $\Psi_0 b = 0$,

$\hat{b}_{\Psi}$ is the mvue of b under the restriction $\Psi b = 0$,

$R^2_{\Psi_0}$ is the squared multiple correlation under the restriction $\Psi_0 b = 0$, and

R^2_{Ψ} is the squared multiple correlation under the restriction $\Psi b = 0$.

The Chipman and Rao computational form for F is different from the form in (7), but the two are equivalent. (See Rao, 1965, pp. 199–200).

MVUE for a Criterion Cluster

The restriction $\alpha_1 = \alpha_2 = \ldots = \alpha_t$ and $b_1 = b_2 = \ldots = b_t$ can be expressed in the form $\Psi b = 0$ as

(8) $$(t-1)(p+1) \left\{ \left[\begin{array}{ccccccccc} I & -I & 0 & \cdots & & & & & \\ \cdot & & -I & 0 & \cdots & & & & \\ \cdot & & & \cdot & \cdot & & 0 & & \\ \cdot & 0 & & & \cdot & & & & \\ & & & \cdot & & 0 & & & \\ I & 0 & \cdots & 0 & -I & 0 & \cdots & 0 \end{array}\right] \right. \begin{bmatrix} \alpha_1 \\ b_1 \\ \vdots \\ \alpha_k \\ b_k \end{bmatrix} = 0$$

$$\underbrace{\qquad\qquad}_{t(p+1)} \quad \underbrace{\qquad\qquad}_{(k-t)(p+1)}$$

where I is the (p+1)x(p+1) identity matrix. To express model (1b) in a form similar to equation (5) under the above restriction (8), the k(p+1)x(k−t+1)(p+1) matrix B, where

$$B' = \begin{array}{c} \overbrace{\qquad t(p+1) \qquad}\ \overbrace{\qquad (k-t)(p+1) \qquad} \\ \left.\left[\begin{array}{cccccccc} I & . & . & . & I & 0 & \ldots & 0 \\ 0 & . & . & . & 0 & I & & \\ & & & & & & \ddots & 0 \\ & & & & & & & \\ & 0 & & & & & & . \\ & & & & & & & I \end{array}\right]\right\} (k-t+1)(p+1) \end{array}$$

and the $(k-t+1)(p+1)$ vector of new parameters θ, where

$\theta' = [\alpha_\Psi\, b_\Psi\, \alpha_{t+1}\, b_{t+1} \ldots \alpha_k\, b_k]$,

is introduced into (1b) to yield the model

$$(9)\quad E(Y|X) = \begin{bmatrix} u_1X_1 & & & 0 \\ \vdots & & & \\ u_tX_t & & & \\ & u_{t+1}X_{t+1} & & \\ & & \ddots & \\ 0 & & & u_kX_k \end{bmatrix} \cdot \begin{bmatrix} \alpha_\Psi \\ b_\Psi \\ \alpha_{t+1} \\ b_{t+1} \\ \vdots \\ \alpha_k \\ b_k \end{bmatrix} \quad \text{with } D(Y|X) = \sigma^2 I.$$

The effect of B is to pool the observations for criteria 1, . . ., t. The mvue $\hat{\alpha}_\Psi$ and $\hat{b}_\Psi$, for the criterion cluster (1, 2, . . ., t) formed from criteria 1, . . ., t can be calculated in either of two ways: pool the observations as in (9) and compute $\hat{\alpha}_\Psi$ and $\hat{b}_\Psi$ from the normal equations

$$(9a)\quad \left\{\begin{bmatrix} n_1 & u_1'X_1 \\ X_1'u_1 & X_1'X_1 \end{bmatrix} + \ldots + \begin{bmatrix} n_t & u_t'X_t \\ X_t'u_t & X_t'X_t \end{bmatrix}\right\} \begin{bmatrix} \hat{\alpha}_\Psi \\ \hat{b}_\Psi \end{bmatrix} = \left\{\begin{bmatrix} u_1'y_1 \\ X_1'y_1 \end{bmatrix} + \ldots + \begin{bmatrix} u_t'y_t \\ X_t'y_t \end{bmatrix}\right\}$$

or if the predictor sums-of-squares and cross-product matrices are proportional, i.e.,

$$(9b)\quad \frac{1}{n_1}\begin{bmatrix} n_1 & u_1'X_1 \\ X_1'u_1 & X_1'X_1 \end{bmatrix} = \frac{1}{n_2}\begin{bmatrix} n_2 & u_2'X_2 \\ X_2'u_2 & X_2'X_2 \end{bmatrix} = \ldots = \frac{1}{n_t}\begin{bmatrix} n_t & u_t'X_t \\ X_t'u_t & X_t'X_t \end{bmatrix},$$

then $\hat{\alpha}_\Psi$ and $\hat{b}_\Psi$ can be calculated from $\hat{\alpha}_1, \hat{b}_1, \ldots, \hat{\alpha}_t$, and $\hat{b}_t$ given in (2) without forming the sum of matrices on the left hand side in (9a). Using (9b) this sum of matrices is

$$(9c)\quad \left\{\begin{bmatrix} n_1 & u_1'X_1 \\ X_1'u_1 & X_1'X_1 \end{bmatrix} + \ldots + \begin{bmatrix} n_t & u_t'X_t \\ X_t'u_t & X_t'X_t \end{bmatrix}\right\} = \frac{N_t}{n_i}\begin{bmatrix} n_i & u_i'X_i \\ X_i'u_i & X_i'X_i \end{bmatrix} \quad \text{for } i = 1, \ldots, t$$

where $N_t = n_1+n_2+ \ldots +n_t$. Using (9c) the solution of (9a) is

$$\begin{bmatrix}\hat{\alpha}_\Psi \\ \hat{b}_\Psi\end{bmatrix} = \sum_{i=1}^{t} \left(\begin{bmatrix} n_1 & u_1'X_1 \\ X_1'u_1 & X_1'X_1 \end{bmatrix} + \ldots + \begin{bmatrix} n_t & u_t'X_t \\ X_t'u_t & X_t'X_t \end{bmatrix} \right)^{-1} \begin{bmatrix} u_i'y_i \\ X_i'y_i \end{bmatrix}$$

$$= \sum_{i=1}^{t} \frac{n_i}{N_t} \begin{bmatrix} n_i & u_i'X_i \\ X_i'u_i & X_i'X_i \end{bmatrix}^{-1} \begin{bmatrix} u_i'y_i \\ X_i'y_i \end{bmatrix}.$$

Thus, the mvue for a criterion cluster are

$$(10) \quad \begin{bmatrix}\hat{\alpha}_\Psi \\ \hat{b}_\Psi\end{bmatrix} = \frac{n_1}{N_t} \begin{bmatrix}\hat{\alpha}_1 \\ \hat{b}_1\end{bmatrix} + \ldots + \frac{n_t}{N_t} \begin{bmatrix}\hat{\alpha}_t \\ \hat{b}_t\end{bmatrix}$$

When (9b) holds, the formula for the standardized regression weights for a criterion cluster is easy to obtain. Let $\hat{\beta}_\Psi, \hat{\beta}_1, \ldots, \hat{\beta}_t$ be the standardized weights corresponding to the raw weights $\hat{b}_\Psi, \hat{b}_1, \ldots, \hat{b}_t$; let Q_i be the pxp diagonal matrix with its elements equal to the standard deviations calculated from the observation matrix X_i for the p independent variables; let Q_Ψ be the pxp diagonal matrix with its elements equal to the standard deviations calculated from the pooled observation matrix $[X_1' X_2' \ldots X_t']'$ for the p independent variables; and let $\sigma_\Psi^2, \sigma_1^2, \ldots, \sigma_t^2$ be the sample variances for the vectors of criterion observations $[y_1' y_2' \ldots y_t']', y_1 \ldots, y_t$, respectively. By definition the standardized weights are

$$\hat{\beta}_\Psi = \frac{Q_\Psi \hat{b}_\Psi}{\sigma_\Psi}, \ \hat{\beta}_1 = \frac{Q_1 \hat{b}_1}{\sigma_1}, \ldots, \beta_t = \frac{Q_t b_t}{\sigma_t}.$$

From (9b), $Q_\Psi = Q_1 = \ldots = Q_t$; therefore using (10), the formula for the standardized weights for a criterion cluster is

$$(10a) \quad \hat{\beta}_\Psi = \frac{1}{N_t \sigma_\Psi} (n_1 \sigma_1 \hat{\beta}_1 + \ldots + n_t \sigma_t \hat{\beta}_t).$$

Multiple Correlation Coefficient for a Criterion Cluster

Let $R_\Psi^2, \hat{R}_1^2 \ldots, \hat{R}_t^2$ be the squared multiple correlation coefficients for the criterion cluster formed from criteria 1, . . ., t and for the t criteria $y_1, \ldots, y_t$, respectively; let c_i be the px1 vector of intercorrelations calculated from the observations X_i and y_i between the p independent variables and the i–th criterion; and let c_Ψ be the px1 vector of intercorrelations calculated from the pooled observations $[X_1' X_2' \ldots X_t']'$ and $[y_1' y_2' \ldots y_t']'$ between the p independent variables and the criterion cluster (1, 2, . . ., t). By definition,

$$n_i \sigma_i Q_i c_i = X_i'y_i - \frac{1}{n_i} X_i'u_i u_i'y_i \qquad \text{for } i=1, \ldots, k \text{ and}$$

$$N_t \sigma_\Psi Q_\Psi c_\Psi = (X_1'y_1 + \ldots + X_t'y_t) - \frac{1}{N_t} [X_1'u_1 + \ldots + X_t'u_t] [u_1'y_1 + \ldots + u_t'y_t].$$

From (9c), $\frac{1}{N_t} [X_1'u_1 + \ldots + X_t'u_t] = \frac{1}{n_i} X_i'u_i$ for $i=1, \ldots, t$. Therefore,

$$N_t \sigma_\Psi Q_\Psi c_\Psi = n_1 \sigma_1 Q_1 c_1 + \ldots + n_t \sigma_t Q_t c_t.$$

But $Q_\Psi = Q_1 = \ldots = Q_t$ so the validity coefficients for a criterion cluster are

$$(10b)\ c_\Psi = \frac{1}{N_t\sigma_\Psi}(n_1\sigma_1 c_1 + \ldots + n_t\sigma_t c_t).$$

The squared multiple correlation coefficient for the cluster

(1, 2, . . ., t) is

$$(10c)\ R_\Psi^2 = \hat{\beta}_\Psi' c_\Psi = \frac{1}{N_t^2\sigma_\Psi^2}(n_1\sigma_1\hat{\beta}_1 + \ldots + n_t\sigma_t\hat{\beta}_t)'(n_1\sigma_1 c_1 + \ldots + n_t\sigma_t c_t).$$

Hypothesis Testing

The critical region given in (7) for the hypothesis (6) requires the calculation of the error sum of squares or the squared multiple correlation coefficient for model (1b) when restrictions are imposed on the unknown parameters. The error sum of squares, ESS, for model (1b) when there are no restrictions on the unknown parameters is equal to the sum of the error sum of squares, ESS_i, for the k models (see (3)), i.e.,

$$ESS = ESS_1 + ESS_2 + \ldots + ESS_k.$$

Let m_0 and σ_0^2 be the criterion mean and variance calculated from the pooled criterion observation vector Y, and let $m_1, \ldots, m_k$ be the criterion means for $y_1, \ldots, y_k$, respectively. Then

$$ESS_i = n_i\sigma_i^2(1 - R_i^2) \quad \text{for } i=1, \ldots, k$$

$$Nm_0 = n_1m_1 + n_2m_2 + \ldots + n_km_k$$

$$N\sigma_0^2 = n_1(\sigma_1^2+m_1^2) + \ldots + n_k(\sigma_k^2+m_k^2) - Nm_0^2$$

Therefore the squared multiple correlation, R^2, for (1b) is

$$(11)\qquad R^2 = \frac{N\sigma_0^2 - ESS}{N\sigma_0^2} = \frac{\left[n_1(\sigma_1^2R_1^2+m_1^2) + \ldots + n_k(\sigma_k^2R_k^2+m_k^2) - Nm_0^2\right]}{\left[n_1(\sigma_1^2+m_1^2) + \ldots + n_k(\sigma_k^2+m_k^2) - Nm_0^2\right]}$$

The error sum of squares, ESSH, for (9) is

$$ESSH = ESS_\Psi + ESS_{t+1} + \ldots + ESS_k$$

where $ESS_\Psi = N_t\sigma_\Psi^2(1 - R_\Psi^2)$. Therefore the squared multiple correlation, R_0^2, for (9) is

$$R_0^2 = \frac{N\sigma_0^2 - ESSH}{N\sigma_0^2}$$

The hypothesis (8) can be tested at the α significance level by computing

$$(11a)\ F = \left(\frac{N - k(p+1)}{(t-1)(p+1)}\right)\left(\frac{R^2 - R_0^2}{1 - R^2}\right)$$

and rejecting (8) if F exceeds the 100(1 - α)% point of the central F distribution with (t - 1)(p+1) and N - k(p+1) degrees of freedom.

Application to a Four Criteria Model; A Worked Example

Given four criteria y_1, y_2, y_3, and y_4, where y_i is an n_ix1 vector of observations, and the predictor matrices X_1, X_2, X_3, and X_4, where X_i is an $n_i$xp matrix of observations on p independent variables, the

greatest predictive power is attained when each criterion variable is predicted from its regression on the independent variables. The initial stage, i.e., Stage 4, employs the following model:

$$(12)\; E\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \underset{4\times 4}{\begin{bmatrix} u_1 X_1 & 0 & 0 & 0 \\ 0 & u_2 X_2 & 0 & 0 \\ 0 & 0 & u_3 X_3 & 0 \\ 0 & 0 & 0 & u_4 X_4 \end{bmatrix}} \underset{4\times 1}{\begin{bmatrix} \alpha_1 \\ b_1 \\ \alpha_2 \\ b_2 \\ \alpha_3 \\ b_3 \\ \alpha_4 \\ b_4 \end{bmatrix}} = \underset{4\times 1}{\begin{bmatrix} \alpha_1 u_1 + b_1 X_1 \\ \alpha_2 u_2 + b_2 X_2 \\ \alpha_3 u_3 + b_3 X_3 \\ \alpha_4 u_4 + b_4 X_4 \end{bmatrix}} \text{ with } D\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \sigma^2 I .$$

The mvue $\hat{\alpha}_i$ and $\hat{b}_i$, for α_i and b_i are obtained from (2) and the squared multiple correlation coefficient, ${}_4R^2$, for model (12) is obtained from (11).

For Stage 3, assume (9b) holds for X_1, X_2, X_3, and X_4. Under the linear hypothesis $\alpha_1 = \alpha_2$ and $b_1 = b_2$, the mvue $\hat{\alpha}_{12}$ and $\hat{b}_{12}$, for the criterion cluster (1,2) formed from criteria 1 and 2 are (see (10))

$$\begin{bmatrix} \hat{\alpha}_{12} \\ \hat{b}_{12} \end{bmatrix} = \frac{n_1}{(n_1+n_2)} \begin{bmatrix} \hat{\alpha}_1 \\ \hat{b}_1 \end{bmatrix} + \frac{n_2}{n_1+n_2} \begin{bmatrix} \hat{\alpha}_2 \\ \hat{b}_2 \end{bmatrix}$$

The standard weights, $\hat{\beta}_{12}$, and the validities, c_{12}, for the cluster (1,2) are (see (10a) and (10b))

$$\hat{\beta}_{12} = \frac{1}{(n_1+n_2)\sigma_{12}} (n_1\sigma_1\hat{\beta}_1 + n_2\sigma_2\hat{\beta}_2), \text{ and}$$

$$c_{12} = \frac{1}{(n_1+n_2)\sigma_{12}} (n_1\sigma_1 c_1 + n_2\sigma_2 c_2), \text{ where}$$

$$(n_1+n_2)\sigma^2_{12} = n_1(\sigma_1^2 + m_1^2) + n_2(\sigma_2^2 + m_2^2) - \frac{(n_1 m_1 + n_2 m_2)^2}{n_1+n_2}$$

The model used to obtain these estimates is (see (9))

$$(13)\; E\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} u_1 X_1 & 0 & 0 \\ u_2 X_2 & & \\ 0 & u_3 X_3 & 0 \\ 0 & 0 & u_4 X_4 \end{bmatrix} \begin{bmatrix} \alpha_{12} \\ b_{12} \\ \alpha_3 \\ b_3 \\ \alpha_4 \\ b_4 \end{bmatrix} = \begin{bmatrix} \alpha_{12}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + b_{12}\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \\ \alpha_3 u_3 + b_3 X_3 \\ \alpha_4 u_4 + b_4 X_4 \end{bmatrix} \text{ with } D\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \sigma^2 I.$$

The squared multiple correlation coefficient, ${}_3R^2$, for (13) is (from (11) with k = 3)

$${}_3R^2 = \frac{\left[(n_1+n_2)(\sigma^2_{12}R^2_{12} + m^2_{12}) + n_3(\sigma^2_3 R^2_3 + m^2_3) + n_4(\sigma^2_4 R^2_4 + m^2_4) - N m_0^2\right]}{\left[(n_1+n_2)(\sigma^2_{12} + m^2_{12}) + n_3(\sigma^2_3 + m^2_3) + n_4(\sigma^2_4 + m^2_4) - N m_0^2\right]}, \text{ where}$$

$$R^2_{12} = \hat{\beta}_{12} c_{12},\quad m_{12} = \frac{(n_1 m_1 + n_2 m_2)}{(n_1+n_2)},\quad N = n_1 + n_2 + n_3 + n_4, \text{ and}$$

$$N m_0 = n_1 m_1 + n_2 m_2 + n_3 m_3 + n_4 m_4$$

(11a) can now be used to test at the α significance level the hypothesis H1: $\alpha_1 = \alpha_2$ and $b_1 = b_2$ by computing

$$F = \left(\frac{N-4(p+1)}{(p+1)}\right) \left(\frac{{}_4R^2 - {}_3R^2}{(1 - {}_4R^2)}\right)$$

and rejecting H1 if F exceeds $F_\alpha(p+1, N-4(p+1))$.

For Stage 2, accepting H1 as true, the additional restrictions $\alpha_3 = \alpha_4$ and $b_3 = b_4$ are imposed and the mvue, $\hat{\alpha}_{34}$ and $\hat{b}_{34}$, for the criterion cluster (3,4) formed from criteria 3 and 4 are

$$\begin{bmatrix} \hat{\alpha}_{34} \\ \hat{b}_{34} \end{bmatrix} = \frac{n_3}{n_3 + n_4} \begin{bmatrix} \hat{\alpha}_3 \\ \hat{b}_3 \end{bmatrix} + \frac{n_4}{n_3 + n_4} \begin{bmatrix} \hat{\alpha}_4 \\ \hat{b}_4 \end{bmatrix}.$$

The standard weights, $\hat{\beta}_{34}$, and the validities, c_{34}, for the cluster (3,4) are

$$\hat{\beta}_{34} = \frac{1}{(n_3 + n_4)\sigma_{34}} (n_3 \sigma_3 \hat{\beta}_3 + n_4 \sigma_4 \hat{\beta}_4), \text{ and}$$

$$c_{34} = \frac{1}{(n_3 + n_4)\sigma_{34}} (n_3 \sigma_3 c_3 + n_4 \sigma_4 c_4), \text{ where}$$

$$(n_3 + n_4)\sigma_{34}^2 = n_3(\sigma_3^2 + m_3^2) + n_4(\sigma_4^2 + m_4^2) - \frac{(n_3 m_3 + n_4 m_4)^2}{(n_3 + n_4)}$$

The model used to obtain these estimates is

$$(14)\ E \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} u_1 X_1 & 0 \\ u_2 X_2 & \\ 0 & u_3 X_3 \\ & u_4 X_4 \end{bmatrix} \begin{bmatrix} \alpha_{12} \\ b_{12} \\ \alpha_{34} \\ b_{34} \end{bmatrix} = \begin{bmatrix} \alpha_{12} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + b_{12} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \\ \alpha_{34} \begin{bmatrix} u_3 \\ u_4 \end{bmatrix} + b_{34} \begin{bmatrix} X_3 \\ X_4 \end{bmatrix} \end{bmatrix} \text{ with } D \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \sigma^2 I.$$

The squared multiple correlation coefficient, ${}_2R^2$, for (14) is (from (11) with k=2)

$${}_2R^2 = \frac{\left[(n_1+n_2)(\sigma_{12}^2 R_{12}^2 + m_{12}^2) + (n_3+n_4)(\sigma_{34}^2 R_{34}^2 + m_{34}^2) - N m_0^2\right]}{\left[(n_1+n_2)(\sigma_{12}^2 + m_{12}^2) + (n_3+n_4)(\sigma_{34}^2 + m_{34}^2) - N m_0^2\right]}$$

where $R_{34}^2 = \hat{\beta}_{34} c_{34}$, $(n_3 + n_4) m_{34} = n_3 m_3 + n_4 m_4$. Equation (11a) can now be used to test at the α significance level the hypothesis

H2: $\alpha_3 = \alpha_4$ and $b_3 = b_4$ given H1 is true by computing

$$F = \left(\frac{N-3(p+1)}{(p+1)}\right) \left(\frac{{}_3R^2 - {}_2R^2}{(1 - {}_3R^2)}\right)$$

and rejecting H2 if F exceeds $F_\alpha(p+1, N-3(p+1))$.

Equation (11a) can also be used to test the hypothesis

H3: $\alpha_1 = \alpha_2$, $b_1 = b_2$, $\alpha_3 = \alpha_4$, and $b_3 = b_4$ by computing

$$F = \left(\frac{N-4(p+1)}{2(p+1)}\right) \left(\frac{{}_4R^2 - {}_2R^2}{(1-{}_4R^2)}\right)$$

and rejecting H3 if F exceeds $F_\alpha(2(p+1), N-4(p+1))$.

For Stage 1, accepting H2 as true, the additional restrictions $\alpha_{12}=\alpha_{34}$ and $b_{12}=b_{34}$ are imposed and the mvue, $\hat{\alpha}_{1234}$ and $\hat{b}_{1234}$, for the criterion cluster (1,2,3,4) formed from all four criteria are

$$\begin{bmatrix} \hat{\alpha}_{1234} \\ \hat{b}_{1234} \end{bmatrix} = \frac{(n_1+n_2)}{N} \begin{bmatrix} \hat{\alpha}_{12} \\ \hat{b}_{12} \end{bmatrix} + \frac{(n_3+n_4)}{N} \begin{bmatrix} \hat{\alpha}_{34} \\ \hat{b}_{34} \end{bmatrix}$$

The standard weights, $\hat{\beta}_{1234}$, and the validities, c_{1234}, for the cluster (1,2,3,4) are

$$\hat{\beta}_{1234} = \frac{1}{N\sigma_{1234}} \left[(n_1+n_2)\sigma_{12}\hat{\beta}_{12} + (n_3+n_4)\sigma_{34}\hat{\beta}_{34}\right], \text{ and}$$

$$c_{1234} = \frac{1}{N\sigma_{1234}} \left[(n_1+n_2)\sigma_{12}c_{12} + (n_3+n_4)\sigma_{34}c_{34}\right], \text{ where}$$

$$N\sigma^2_{1234} = (n_1+n_3)(\sigma^2_{12}+m^2_{12}) + (n_3+n_4)(\sigma^2_{34}+m^2_{34}) - Nm^2_{1234}, \text{ and}$$

$$Nm_{1234} = n_1m_1+n_2m_2+n_3m_3+n_4m_4 .$$

The model used to obtain the estimates for cluster (1,2,3,4) is

$$(15)\quad E\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} \begin{bmatrix} u_1 & X_1 \\ u_2 & X_2 \\ u_3 & X_3 \\ u_4 & X_4 \end{bmatrix} \begin{bmatrix} \alpha_{1234} \\ b_{1234} \end{bmatrix} = \left[\alpha_{1234}\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} + b_{1234}\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix}\right] \text{ with } D\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \sigma^2 I.$$

The squared multiple correlation coefficient, ${}_1R^2$, for (15) is

$${}_1R^2 = \hat{\beta}'_{1234}\, c_{1234}.$$

Equation (11a) can now be used to test at the α significance level the hypothesis

H4 : $\alpha_{12}=\alpha_{34}$ and $b_{12}=b_{34}$, given $\alpha_1=\alpha_2$, $\alpha_3=\alpha_4$, $b_1=b_2$ and $b_3=b_4$ by computing

$$F = \left(\frac{N-2(p+1)}{(p+1)}\right) \left(\frac{{}_2R^2 - {}_1R^2}{(1-{}_2R^2)}\right)$$

and rejecting H4 if F exceeds $F_\alpha(p+1, N-2(p+1))$. The hypothesis

H5 : $\alpha_1=\alpha_2=\alpha_3=\alpha_4$ and $b_1=b_2=b_3=b_4$

can be tested at the α significance level by computing

$$F = \left(\frac{N-4(p+1)}{3(p+1)}\right) \left(\frac{{}_4R^2 - {}_1R^2}{(1-{}_4R^2)}\right)$$

and rejecting H5 if F exceeds $F_\alpha(3(p+1), N-4(p+1))$.